Homework 1: Random Events and Probability

UA CSC 380: Principles of Data Science

Homework due at 11:59pm on Jan 27, 2023

Deliverables Your must make two submissions: (1) your homework as a SINGLE PDF file by the stated deadline to the gradescope. Include your code and output of the code as texts in the PDF. and (2) your codes as a ZIP file to a separate submission. Each subproblem is worth 10 points. More instructions:

- You can hand-write your answers and scan them to make it a PDF. If you use your phone camera, I recommend using TurboScan (smartphone app) or similar ones to avoid uploading a slanted image or showing the background. Make sure you rotate it correctly.
- Watch the video and follow the instruction for the submission: https://youtu.be/KMPoby5g_nE
- **Show all work along with answers to get the full credit**.
- Place your final answer into an 'answer box' that can be easily identified.
- There will be no late days. Late homeworks result in zero credit.

Failure to follow the submission instruction will result in a minor penalty in credit.

You can choose to work individually or in pairs.

- If you choose to work in pairs, you are free to discuss whatever you want with your partner; please make only one submission per group.
- Please do not discuss with people outside your group about the homework (refer to the academic integrity policy in Lecture 1).
- If you have clarification questions, please feel free to post on Piazza so that it can promote discussion.

Problem 1: Random Dice

This problem will compare the theoretical properties of a fair die to empirical results from simulation. It will further familiarize you with the `numpy.random` library.

a) Assume that we roll two fair six-sided dice. Let E be the event that the two dice's outcomes sum to 3. What is the probability of E?

b) Initialize the random seed to 2023 using `numpy.random.seed`. Using `numpy.random.randint`, simulate 1,000 throws of two fair six-sided dice. Paste your code here. From these simulations, what is the empirical frequency of E (i.e., the percentage of times this event occurred in simulation)?

c) Reset the random seed to 2023 and repeat the above simulation a total of 10 times and report the empirical frequency of E for each run. Paste your code here.

d) The empirical frequency of E from each simulation will differ. Why do these numbers differ? Yet, the probability of E is fixed and was calculated in part (a) above. Why does the probability disagree with the empirical frequencies?

e) In the above we have estimated the probability of an event by performing $1,000$ rolls of two dice each. We generated 10 different estimates by repeating this procedure. How do our results change if we instead perform $10,000$ rolls, and repeat 10 times? Try it, report the difference, and discuss why. Paste your code here.

Problem 2: Coinflips

Suppose we flip a fair coin 10 times. What is the probability that the following events occur: I recommend that you use the code like Problem 1 to debug your answers (but this debugging itself is not part of the evaluation).

a) The number of heads and the number of tails are equal

b) There are strictly more heads than tails

c) The number of heads and the number of tails are equal, but now with the assumption that the head probability is .3 (unfair coin).

Problem 3: Conditional Probability

a) Assume that we roll two fair six-sided dice. What is $P(\text{sum is } 5 \mid \text{first die is } 2)$? What is $P(\text{sum is } 5 \mid \text{first die is } 5)$?

b) Assume that we roll two fair four-sided dice. What is $P(\text{sum is at least } 4)$? What is $P(\text{First die is } 1)$? What is $P(\text{sum is at least } 4 \mid \text{first die is } 1)$?

c) Suppose two players each roll a die, and the one with the highest roll wins. Each roll is considered a "round" and further suppose that ties magically don't happen (or those rounds are simply ignored) so there is always a winner. The best out of 7 rounds wins the match (in other words the first to win 4 rounds wins the match). Let W be the event that you win the whole match. Let $S = (i, j)$ be the current score where you have i wins and the opponent has j wins. Compute the probability that you win the match, given the current score, i.e., $a_{i,j} := P(W \mid S = (i, j))$ for each of the 16 possible values of $S = (i, j)$, $i, j \in \{0, 1, 2, 3\}$.

To help you out a bit with part (c) above, I am providing the following hints:

1) If $i = 4$ and $j < 4$, $a_{i,j} = 1$. OTOH, if $i < 4$ and $j = 4$, $a_{i,j} = 0$. Can you see why?

2) Let R_i be a random variable where $R_i = 1$ if you win round i and $R_i = 0$ if you lose that round. Note that $P(R_i = 0) = P(R_i = 1) = \frac{1}{2}$.

3) Recall that by the law of total probability $P(W \mid S = (i, j)) = P(W, R_{i+j+1} = 1 \mid S = (i, j)) + P(W, R_{i+j+1} = 0 \mid S = (i, j))$.

4) By the probability chain rule $P(W, R_{i+j+1} \mid S = (i, j)) = P(W \mid R_{i+j+1}, S = (i, j))P(R_{i+j+1} \mid S = (i, j))$.

5) Although it requires rigorous argument, for this problem, you can take it as given that $P(W \mid R_{i+j+1} = 1, S = (i, j)) = P(W \mid S = (i + 1, j))$ and $P(W \mid R_{i+j+1} = 0, S = (i, j)) = P(W \mid S = (i, j + 1))$. (Can you see why, intuitively?)

6) Find a way to write down $a_{i,j}$ as a function of $a_{i+1,j}$ and $a_{i,j+1}$. This will help you compute the answers in a recursive manner.

7) As a sanity check, when the current score is equal–that is $S = (k, k)$–then there should be equal chance of either player winning the match, $P(W \mid S = (k, k)) = \frac{1}{2}$ for $k = 0, 1, 2, 3$.